F-052

単語の重要度と頻度を利用した局所探索に基づく WEB ページ発見法
Finding Web Pages Using Local Search with Word Importance and Frequency

上條 朋彦† 栗原 正仁‡
Tomohiko Kamijyo Masahito Kurihara

1. はじめに

今日，インターネットは急速な発展をみせており，WEB ページの数は日々増加し続け，その総数は数十億を超えたとされている．このような状況において，手動で全てのページを探索していくのは非現実的であり，WEB から有益な情報を効率的に取り出す方法が求められる．

現在，インターネットユーザの多くが利用している検索エンジンでは，PageRank[1]や HITS[2]アルゴリズムがページ解析手法として用いられている．これらの手法では WEB 構造全体，大規模な WEB ページ集合から各ページを評価している．しかし，閲覧ユーザにとっては，より有益なページが既存手法で得たページの周辺に存在している場合も考えられ，閲覧ユーザの意図に沿ったページ集合を取得した上で，重点的に解析を行う手法が無いのが現状である．また，既存手法では WEB ページのリンク構造・内容を示すテキスト・アクセスログを利用した解析を行うが，これらの情報は主にページ作成者や他の閲覧ユーザの意図を反映した情報であり，今現在ページを探しているユーザの意図は，これらの情報には含まれていない．

本論文では，既存の Web 解析手法の問題点を解決・緩和するため，よりユーザの目的に合った関連ページの発見手法を提案する．本手法は，HITS をユーザの評価をより反映させるように応用したものである．また，ページ集合の取得時にもユーザの意図を反映させている．

さらに，計算機実験では検索エンジン Google の補助に重点を置いた実験を行い，その結果から既存手法では発見できなかったページ発見の可能性について示す．

2. 提案手法

本手法の基本的なアルゴリズムは HITS[2]を応用したものであり，おおまかな流れは次の 3step で構成される．

step1. ベースページ集合の取得とユーザによる評価
step2. 評価値を利用したページ収集
step3. 収集ページを元に，評価値を利用して解析

以後，各 step について詳述していく．

2.1 ベースページの取得と評価

ページ探索の始点となるベースページ集合は，ユーザの目的に近いページを直接指定するか，あるいは HITS と同様に検索エンジンにおいてトピックワードの検索結果上位一定件数を利用する等の方法で取得する．

得られたベースページ集合の各ページのタイトルおよび本文から，単語を形態素解析法により抽出を行う．このとき抽出する品詞は，重要語となりやすい名詞・形容詞・形容動詞に限る．

形態素解析によって得られた単語に対してユーザにより評価値を与える．単語 i 評価値 w_i は 0~5 の数値で表現する．

$$w_i = [0..5] \qquad \text{-(1)}$$

2.2 単語の評価値を利用したページ収集

ページ収集を行う前に，評価値を元に各ベースページに対してページ評価値 SCORE を与える．SCORE は以下の式で表される．

$$SCORE(p) = \sum_{i \in Q} w_i \times freq_{pi} \qquad \text{-(2)}$$

SCORE は，ユーザが重要視した単語が数多く含まれるページの評価を高くする．この SCORE の値が最も高いページからハイパーリンクを局所的にたどりながら，ページ集合を取得していく．具体的には，ページをノードとし，それらをつなぐハイパーリンクを有向枝とする探索木を，SCORE が最大のページを根ノードとし，木の先端ノードの SCORE を評価関数とするヒューリスティック探索（最良優先探索）によって成長させていき，探索木に含まれるノード集合を収集する．

ページ集合の打ち止め条件は，場合によって様々な指定方法が考えられるが，本論文では比較検証のため，指定したページ数の上限を超えるまでとした．

2.3 評価値を利用したページ解析

最後に，収集したページに対し，本手法に応用させた HITS[2]アルゴリズムにより解析を行い，その結果をユーザにとって有用なページとして提示する．

HITS では，Web ページを Authority ページと Hub ページの 2 種類に分類し，解析を行う．ここで，Authority ページとはリンクを数多く張られている公式ページ群のようなものであり，Hub ページは，優秀なリンクを多数発しているリンク集のようなページを指す．この 2 種類の観点から上位となるページを，推薦ページとしてユーザに提示する．

提案手法では，Authority および Hub の初期値に SCORE を用いることで，HITS を本研究に応用させて適用する．

†‡北海道大学大学院 情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

3. Google 検索補助実験

3.1 実験設定

実験設定では, ベースページを検索エンジン Google[3] の検索結果の上位の中からユーザが選択し, ベースページ集合とする.

状況としては, ユーザがファイル解凍ソフトの入手を試みている場合を想定し, 検索ワードには「ダウンロード ソフトウェア 解凍」という文字列を設定する. このキーワードによる検索結果から選択したページは, 表 1. に示す 3 つで, 検索結果の順位 1, 4, 6 位のページとする. また, 抽出された単語に対して与える評価値は表 2. の通りとする.

表 1. ベースページ集合

URL
http://www.nifty.com/download/begin.htm
http://www.forest.impress.co.jp/
http://www.jp.sonystyle.com/peg/Store/Software/Guide/

表 2. 単語に与える評価値

単語	評価値
ダウンロード	5
圧縮	3
解凍	5
おすすめ	3
ソフトウェア	3

3.2 実験結果と考察

上記の実験設定において, 収集する Web ページのハイパーリンク数を 300 件, 1500 件とした場合の実験結果をそれぞれ表 3, 4 および表 5, 6 に示す. どちらの場合も, Vector(http://www.vector.co.jp/)およびその内部ページが上位を占めている. Vector は Google の検索結果には含まれていなかったページであり, Vector の PageRank は 7 である. この値は検索結果上位 10 件のどのページよりも高くなっており, より有益なページが発見できた事を示す.

表中の No は, ページ探索中の発見した順番を示しているが, 最大でも 200 台のものしか出現していない. すなわち, 解析結果には収集件数の差はほぼ無いと言って良い. これは, Vector を早期発見したためであり, 本手法では全てのリンクを辿った場合と比べて有益なページを効率的に発見することが可能である事を示している.

Vector 付近のページのみが推薦されているのは, SCORE 値の高いページが Vector の内部サイトに集中していたため, 優先的に探索された結果である. また, Vector 内部のサイト構造形成のためのハイパーリンクの影響も受けている. これは HITS の特性(トピックドリフト問題)の影響もあるが, 提案手法では HITS の初期値に SCORE を導入しているため, この問題を緩和する可能性がある. また, SCORE 値を利用したフィルタリングを用いることで, 解析結果から無関係なページを除外する事も可能である.

本手法では HITS アルゴリズムを応用しているので, ユーザの目的ページが Authority・Hub の概念と適合する場合は, ユーザの目的に近いページが出現する可能性は高い. それ以外の場合でも, HITS は一般に著名なサイトを推薦するため, 一定の再現率は期待できる.

表 3. Authority 上位 5 件(300 件収集時)

Rank	URL	No	SCORE
1	http://passport.vector.co.jp/	90	0
2	http://maglog.jp/	98	0
3	http://www.galge.com/	96	0
4	http://www.valumore.jp/shop/	95	20
5	http://www.vector.co.jp/games/	92	18

表 4. Hub 上位 5 件(300 件収集時)

Rank	URL	No	SCORE
1	http://www.vector.co.jp/	13	122
2	http://shop.vector.co.jp/service/...	200	0
3	http://shop.vector.co.jp/service/...	201	0
4	http://shop.vector.co.jp/service/...	202	0
5	http://shop.vector.co.jp/service/...	203	0

表 5. Authority 上位 5 件(1500 件収集時)

Rank	URL	No	SCORE
1	http://passport.vector.co.jp/	90	0
2	http://maglog.jp/	98	0
3	http://www.galge.com/	96	0
4	http://www.valumore.jp/shop/	95	20
5	http://security.vector.co.jp/serv...	91	0

表 6. Hub 上位 5 件(1500 件収集時)

Rank	URL	No	SCORE
1	http://www.vector.co.jp/	13	122
2	http://shop.vector.co.jp/service/...	200	0
3	http://shop.vector.co.jp/service/...	201	0
4	http://shop.vector.co.jp/service/...	202	0
5	http://shop.vector.co.jp/service/...	203	0

4. おわりに

本研究では, HITS アルゴリズムを応用し, 閲覧側ユーザの意図を反映させた局所的なページ探索を利用した, 閲覧ユーザの目的に合ったページの発見手法を提案した.

さらに Google 検索エンジン補助実験を行い, 検索エンジンには表示されなかったユーザにとって有益なページが発見可能である事を確認した.

本研究の今後の課題としては, 応用 HITS アルゴリズムのさらなる検証と改善, 単語の評価値およびページ収集打ち止め条件の自動決定化がある.

参考文献

[1]Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd, "The pagerank citation ranking: Bringing order to the web.", Technical report, Stanford Digital Library Technologies Project, (1998).

[2]Jon. M. Kleinberg, "Authoritative sources in a hyperlinked environment.", Journal of the ACM, Vol.46, No.5 , (1999).

[3]http://www.google.co.jp/